Tagging and Glossing Sesotho

Mark Johnson, Katherine Demuth and Stephen Canon
Cognitive and Linguistic Sciences, Box 1978
Brown University
Providence, RI 02912
Mark_Johnson@Brown.edu
Katherine_Demuth@Brown.edu
Stephen_Canon@Brown.edu

December, 1999

Abstract

This paper describes a system for morphological tagging and glossing of Sesotho, a southern Bantu language. Sesotho has a rich agglutinative morphology, and morphemes cannot be disambiguated on the basis of the bigram or trigram statistics that work so well for languages like English. Our system estimates a simple PCFG for Sesotho clauses from a small hand-annotated corpus in an unsupervised manner. It uses this PCFG and a small set of hand-coded constraints to produce a ranked list of possible tags and corresponding glosses for untagged clauses.

1 Introduction

This paper describes a system for morphological tagging and glossing of Sesotho, a Bantu language spoken in South Africa and Lesotho. One of us has collected a substantial corpus of spoken Sesotho child and child-directed adult speech. The corpus consists of 59,963 clauses containing 108,580 morphologically complex words. As discussed below, Sesotho is a richly agglutinative language with extensive "pronoun-drop", so a single word may incorporate subject and object pronouns, and verbal inflection and derivation as well as the verb root itself; i.e., what would be 6 or 7 separate words in an isolating language such as English. Thus the number of clauses, rather than the number of words, is probably a better indicator of the corpus' size.

Transcribing the Sesotho recordings into standard Sesotho is a relatively straight-forward and quick task for native speakers. Because the corpus contains informal connected speech in which sentence boundaries are sometimes unclear we decided to transcribe it in clausal units, and used a special annotation to indicate that the current and previous clauses belong to the same sentence.

The resulting raw Sesotho text corpus is not that useful on its own. Tagging this corpus enables us to search it for interesting linguistic constructions. Glossing the corpus enables non-native speakers to understand the gist of the Sesotho text: we wanted to do this to enable students to work with this corpus. Our tagging and glossing conventions are discussed in detail in section 2.

Manually annotating a corpus of this size would entail a substantial effort. Whereas literate Sesotho speakers can transcribe quicky and accurately, manual annotation is slow, tedious and error-prone, and requires bilingual native speakers of Sesotho with linguistic training who are difficult to find outside of southern Africa. Thus we decided to develop the methods described in this paper.

Our goal is to annotate the corpus with a minimum of effort as easily and as accurately as possible using the human resources we have available. We used standard Lesotho orthography because it can be transcribed quickly and accurately, while in our experience additional linguistic transcription (phonetic, phonological or morphological) is slow and error-prone. In terms of implementation, we wanted a system which was both easy to build and to maintain. This ruled out approaches which require complex hand-built dictionaries and rule systems (which are often incomprehensible to anyone but their original authors). Our method, which can be regarded as an extension of HMM techniques to an agglutinative language, requires only a modest tagged and glossed training corpus, from which the required rules and dictionaries are automatically extracted. Extending the system primarily involves adding additional tagged and glossed examples to the training corpus and recomputing the rules and dictionaries. While these rules do not incorporate all possible linguistic constraints, the simplicity, clarity and ease of maintainance of our system more than makes up for the extra ambiguities it returns. Moreover, the rule system is stochastic, which permitted us to return a ranked list of possible analyses for each sentence. As we discuss in section 5, in most cases the first analysis is correct, which eases the manual correction task.

Thus the primary objective of this project was to prepare the corpus for further scientific research. But as an unexpected bonus, we found that the dictionaries and the PCFG we extracted from our corpora in order to build this system were also useful for quality control: any clauses from our corpus from which low count lexical entries or PCFG rules were extracted were also tagged for manual checking. A large number of these clauses did in fact contain typographical errors (which were then corrected), which suggests that the techniques we developed could be used in spell-checking. Indeed, it seems that tagging and glossing are important techniques for many applications of computational linguistics to agglutinative languages, such as document indexing, and they may possibly provide the basis for more advanced applications such as machine translation.

2 A brief description of Sesotho morphology

Sesotho is an agglutinating language, which means that its words are formed by concatenating morphemes according to a rich set of derivational and inflectional morphological rules. For example, a noun such as *motho* ‘person’ consists of a noun class prefix *mo-* and a noun stem *-tho.* The noun class prefix indicates whether the noun is singular or plural (c.f., *batho* ‘people’), and agrees with the noun. Sesotho has 12 distinct noun classes. The noun class prefix for a noun stem depends on the stem itself, and whether the noun is singular or plural. The noun class system can be viewed as a kind of rich gender agreement system. The class that a noun belongs to is essentially arbitrary, although there are both phonological and semantic subregularities that we ignore. The class of the singular form of a noun determines the class of its plural and vice-versa, so count noun stems can be viewed as belonging to two classes; the actual class being selected on the basis of the noun’s number. We infer the classes of noun stems from our training data. Our training data contains morphologically segmented nouns such as *mo-tho* annotated ‘nˆ1-person’. From this we infer that *mo-* is a noun class marker for class 1 noun stems and that *-tho* is a stem whose singular belongs to class 1 and whose plural belongs to class 2 (we represent this by assigning it the annotation ‘person(1,2)’ in our dictionary).

Verbal morphology is more complex. A typical inflected verb has a morphological structure shown in (1), where SM is a subject marker, TA is a tense/aspect marker, OM is an object marker, VROOT is the verbal root, VDERIV are derivational morphemes (there is a fixed set of these, which includes passive, causitive, applicative, etc.), and MOOD is a final mood morpheme (which marks the indicative, imperative and subjunctive, amongst other things).

$$\textsc{sm} - \textsc{ta} - (\textsc{om}) - \textrm{V}\textsc{root} - \textrm{V}\textsc{suffix}^{\star} - \textsc{mood} \qquad (1)$$

The subject and object markers agree in class with the subject and object respectively. The subject marker is obligatory except for imperatives, while the object marker is only required when the verb is transitive and the object NP is dropped (see Demuth and Johnson (1989) for further details). While there are complex restrictions on the number and order of the verbal derivational suffixes, between zero and two of these usually appear on a verb.

Morpheme tags identify not only the type of morpheme (e.g., subject marker, verbal derivational suffix, etc.) but also include an annotation as to the class involved. For example, *ke-* is tagged 'sm1s-', indicating that it is the subject marker agreeing with first person singular subjects. In addition, in order to help non-native speakers understand the morphological structure we incorporate an English gloss of the open-class morphemes in our annotation. For example, the imperative *bon-a* is annotated 'v^see-m^i', indicating that *bon-* is the verb root "see" and *-a* is the imperative singular mood morpheme (thus *bona* means "Look!").

Our system's task is to identify these annotations from Lesotho orthographic forms. Morphological segmentation is itself a non-trivial problem. The task is further complicated by the fact that in standard Lesotho orthography the subject marker and tense/aspect marker are separated from the rest of verb by spaces, as shown in (2). Thus spaces cannot be interpreted as unambiguous word boundaries.

(2) Orthographic form: ke a ba bona batho
Morphologized form: ke-a-ba-bon-a ba-tho
Morpheme gloss: sm1s-t^p-om2-v^see-m^in n^2-person(1,2)
English gloss: "I see them, the people"

3 Standard approaches to morphological tagging and glossing

This section briefly reviews the standard approaches to morphological tagging and glossing, and explains why they are not directly applicable to our task.

3.1 Finite-state transducers

Since Koskenniemi's (1983) pioneeriong work, phonological and morphological analysis has been a classic application for finite-state transducer techniques. These techniques are particularly well-suited to languages with complex phonological rule systems (Kaplan and Kay, 1994). However Sesotho, like most Bantu languages, has a particularly simple phonology; the orthographic constraints required in our system are implemented via simple string substitution. Thus one of the major advantages of finite-state techniques does not apply here.[1] We found that the freely available finite state morphological analysis packages required relatively complex hand-built rule systems and dictionaries, which we felt would be difficult to maintain and update. Further, we wanted a system which would return a ranked set of possible analyses, as this simplifies the manual disambiguation of the system's ambiguous outputs, which standard systems do not provide.

3.2 HMM tagging

Currently, the standard methods for part of speech tagging for isolating languages such as English are statistically based (Manning and Schütze, 1999). One of the simpler but effective algorithms treats the part of speech tags as the hidden states of an HMM generating the word string. In principle these techniques should also be applicable to morphological tagging, where we take the input to be a sequence of morphemes rather than words.

In fact, our system is a direct attempt to extend the standard HMM tagging technology to deal with an agglutinative language like Sesotho. One problem is that we wish our annotation to include phonologically null morphemes. For example, there is no phonologically realized tense/aspect morpheme in (3), but the example is interpreted as present tense, and we wish

[1] Sesotho is a grammatical tone language, and the tone rules are quite complex. Lesotho orthography does not transcribe tone, so we ignore tone here.

to annotate it as such.[2] We describe how we deal with these null morphemes in section 4.

(3)			
	Orthographic form:	ke batla dijo	
	Morphologized form:	ke-∅-batl-a	di-jo
	Morpheme gloss:	sm1s-tˆp-vˆwant-mˆin	nˆ8-food(7,8)
	English gloss:	"I want food"	

However, some of the independence assumptions standardly made in HMM tagging do not hold for sequences of Sesotho morphemes. The PCFG model presented in section 4 remedies this.

In HMM tagging, Bayes inversion is used to relate the probability $\Pr(t_{1,n}|w_{1,n})$ of a tag sequence $t_{1,n}$ given a word sequence $w_{1,n}$ to the probability of the words given the tags and the probability of the tags alone.

$$\begin{aligned} \Pr(t_{1,n}|w_{1,n}) &\propto \Pr(w_{1,n}|t_{1,n})\,\Pr(t_{1,n}) && (4) \\ &= \Pr(w_{1,n}|t_{1,n}) \left(\prod_{i=1}^{n} \Pr(t_i|t_{1,i-1}) \right) && (5) \end{aligned}$$

Several independence assumptions are usually made in order to estimate the quantities in the right hand side of (5). A standard second-order Markov assumption is that the conditional probability of the tag t_i of the ith word is independent of the preceeding tags $t_{1,i-3}$ given the two preceeding tags $t_{i-2,i-1}$, i.e.:

$$\Pr(t_i|t_{1,i-1}) \approx \Pr(t_i|t_{i-2}, t_{i-1}) \qquad (6)$$

Unfortunately, Sesotho morphology exhibits "long-distance" dependencies that a simple tri-tag model based on (6) cannot capture. We give a simple example here involving the verb root *tl*- 'vˆcome' and the causative morpheme -*is*-, annotated 'c', which together mean "bring". The final mood morpheme indicates (among other things) indicative versus imperative plural mood, as in (7a) and (7b) (where it is tagged 'mˆin' and 'mˆip' respectively).

(7) a. Orthographic form: ke a e tlisa
Morphologized form: ke-a-e-tl-is-a
Morpheme gloss: sm1s-tˆp-om9-vˆcome-c-mˆin
English gloss: "I am bringing it"

[2] We invented a special annotation for these null morphemes in order to ease their transcription, which we do not describe here for reasons of space.

b. Orthographic form: tlisang
Morphologized form: tl-is-ang
Morpheme gloss: vˆcome-c-mˆip
English gloss: “Bring (it) (pl.)!”

However, the singular imperative form of the final mood morpheme is homophonous with the singular indicative form, as shown in (8a) where it is tagged ‘mˆi’.

(8) a. Orthographic form: tlisa
Morphologized form: tl-is-a
Morpheme gloss: vˆcome-c-mˆi
English gloss: “Bring (it)!”

Note that a three-tag window could not disambiguate the final mood morpheme -*a* in examples (7a) and (8a), as the preceding two morphemes are identical in both cases. Furthermore, using the productive verbal derivational morphology discussed in section 2, longer verbal chains can be produced. The key point is that with a complex derivational structure, the information required to disambiguate the final mood morpheme will not be available in a two-word window. This motivates the PCFG-based model that we used in our system.

4 Word and clause structure rules

Our system can be viewed as an attempt to generalize the HMM part of speech tagging system just described in two ways. First, we wanted to capture longer-distance morphological dependencies that a fixed-length window would miss. Second, if we were to treat our tag/gloss annotations as part of speech tags directly (i.e., if we treated the English gloss as part of the tag) then the number of possible tags would be close to the size of the lexicon, and we would have serious sparse data problems.

While we experimented with a number of different extensions to the basic HMM approach, we found that the PCFG approach provides a clear and theoretically well-motivated framework in which to formulate and explore alternatives. The PCFG rules we use are instances of the schemata shown in (9-12), which we explain below.

$$A \;\to\; M \tag{9}$$

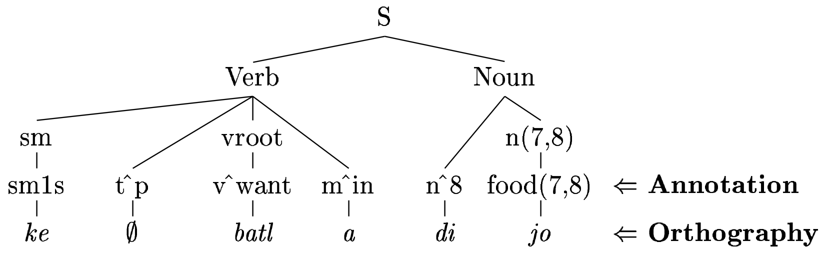

Figure 1: The parse of example 3 using our PCFG.

$$T \rightarrow A \tag{10}$$
$$W \rightarrow (T|A)^\star \tag{11}$$
$$\mathrm{S} \rightarrow W^\star \tag{12}$$

In these schemata, M ranges over morphemes (e.g., *bon*), A ranges over annotations (e.g., 'v^see'), T ranges over morphological tags (e.g., 'vroot'), W ranges over word categories (e.g., Verb), and S is the start symbol. Figure 1 shows a parse of example 3 using our grammar.

As Manning and Schütze (1999) explain, the PCFG rule $t_i \rightarrow w_i$ expresses the same independence assumptions as the HMM approximation that $\Pr(w_i|t_{1,n}) \approx \Pr(w_i|t_i)$. Inspired by this, in our system each morpheme w is introduced by a PCFG rule of the form $A \rightarrow w$, where A is a possible annotation of w, so each morpheme's annotation can be read directly from the parse. Because our annotations are richer than standard HMM tags, we introduced additional unary rules which map more abstract part of speech tags to the annotations we used for open-class morphemes, in effect splitting the HMM approximation into two parts for these morphemes. Thus the rules for expanding the annotations for open-class items are essentially deterministic (e.g., the only rule expanding 'v^want' is 'v^want → *batl*'). We decided to encode the class of the nouns at the tag level, as some noun class markers are phonologically null and others are homophonous with other morphemes. We also decided to ignore the agreement classes of subject and object markers, i.e., there are rules which map the more abstract tags 'sm' and 'om' to each possible subject and object marker annotation respectively.

In order to capture the long-distance dependencies mentioned in section 3.2 we introduced phrase structure rules to generate the morpheme tag sequences found in any word in our training data (these rules were collected automatically; more on this below). There are 37 such rules expanding 'Noun' and 380 rules expanding 'Verb', a not unmanagable number.[3] These word-level rules are particularly important, as they capture linguistically important dependencies, e.g., between the form of the tense/aspect marker and the final mood morpheme.

Our word-level phrase-structure rules also incorporate information as to which morphemes are separated orthographically by spaces: in effect, we introduce dummy categories which rewrite as orthographic spaces, which we ignore when reading off the morpheme annotations.

We initially experimented with a simple model which permits a clause to consist of any sequence of words, but we found that this permitted too many spurious ambiguities. For example, the orthographic form *ke* is ambiguous; it can either be the 1st person singular subject marker, as in (13a) or the 3rd person singular copula, annotated 'cp' in (13b). But notice then that if each word in the clause were assumed to be independent the remaining verbal morphemes in (13a) could be incorrectly analyzed as an imperative verb as in (13c); i.e., (13a) would receive the incorrect annotation 'cp vˆsee-mˆi nˆ9-doctor(9,10)'.

(13) a. Orthographic form: ke bona ngaka
Morphologized form: ke-∅-bon-a ∅-ngaka
Morpheme gloss: sm1s-tˆp-vˆsee-mˆin nˆ9-doctor(9,10)
English gloss: "I see the doctor"

b. Orthographic form: ke ngaka
Morphologized form: ke ∅-ngaka
Morpheme gloss: cp nˆ9-doctor(9,10)
English gloss: "S/he is a doctor"

c. Orthographic form: bona
Morphologized form: bon-a
Morpheme gloss: vˆsee-mˆi
English gloss: "Look!"

In order to force the correct disambiguation of examples like (13a) we introduced clause-level phrase-structure rules, which specify the permissible sequences of word categories. (The incorrect analysis of (13a) is ruled out

[3] This number could have been dramatically reduced by introducing an recursive 'vstem' category, which expands to a verb root followed by zero or more derivational morphemes.

by treating copulas and verbs as a distinct word level categories). Since our corpus was broken into clause-level chunks during transcription clause-level rules seem particularly natural, but we believe this approach could have been used to analyse sentential units as sequences of clauses. Our PCFG contains 1,400 sentence-level rules.

The phrase structure trees used in our system are simple enough that they can be computed deterministically from the annotated orthographic forms in our training corpus. (Recall that our annotations indicate the location of word-boundaries, even though the orthography does not). We needed only a handful of simple deterministic rules to map annotations to the corresponding tags where appropriate, and to identify the word-level categories of each word. In effect we constructed a crude tree-bank for our training corpus, and extracted the required rules and estimated their probabilities from this treebank.

Annotation and segmentation of unannotated orthographic forms merely requires parsing these forms with our PCFG. We used a best-first top-down parser to do this, which handles the job of predicting empty morphemes in a simple, elegant way.

A major component of HMM taggers is usually concerned with handling unknown words. We do not attempt to process clauses containing unknown morphemes, but simply flag the clause for human annotation. This is because most unknown morphemes are open class morphemes, for which we want provide an English gloss. While it is plausible to imagine that the HMM techniques for guessing the tags of unknown words could be used to guess the tags for unknown morphemes here, there seems to be no way to guess the English gloss (short of exploiting a machine-readable dictionary—currently unavailable for Sesotho).

5 Evaluation

Our corpus contains 59,963 clauses consisting of 189,102 (orthographic) words. Of these 11,222 clauses (27,401 words) were manually annotated. From these we held out 1,300 clauses as a test corpus, and extracted a PCFG as described above from the remaining annotated examples. We then parsed the held-out test corpus with the PCFG, and compared the maximum likelihood parses with the manual annotations. Because of sparsity of the training data, the PCFG does not generate every orthographic form in the test corpus. We calculated precision and recall measures for our system. The precision is the ratio of the number of correctly assigned morpheme annotations to the total number of morphemes annotated; it measures the reliability of the annotations that were actually assigned. The recall is the ratio of the number of correctly assigned morpheme annotations to the total number of morphemes; it measures how likely it is that a morpheme will in fact receive the correct annotation. On our test corpus our system achieves a precision of 95.6%, and a recall of 68.5%. As discussed above, the gap between precision and recall seems to be largely due to open-class lexical items.

6 Conclusion

This paper has described a useful system for annotating Sesotho text. It has the advantage of simplicity; even the algorithmically illiterate find it easy to understand why the system behaves as it does, and perhaps more importantly, are able to extend it without expert assistance simply by providing it with additional annotated training examples. It is currently being used to help annotate the remaining portion of our corpus.

The system could undoubtedly be extended in several ways. For example, the system currently flags the entire clause containing any unknown morphemes for manual annotation, even though usually only one morpheme is unknown. It should be possible to isolate and identify the unknown morpheme's orthographic form with only a fairly modest change to our system, e.g., by incorporating a simple PCFG model of the orthographies of "unknown" morphemes. This would both raise our precision score and simplify the manual annotation effort, albeit at the expense of complicating the annotator's task.

Finally, Sesotho has linguistic dependencies which our simple PCFG cannot capture, but which might be useful for disambiguating morpheme annotations. For example, WH-dependencies and relative clauses trigger special kinds of final mood morphemes, which our system currently ignores. One can in fact envisage a system capable of exploiting these dependencies—they could be captured using a version of GPSG's "slashed categories", for example—but at the cost of giving up the conceptual and implementational simplicity of the present system.

References

Katherine Demuth and Mark Johnson. 1989. Interactions between discourse functions and agreement in Setawana. *Journal of African Languages and Linguistics*, 11(21-35).

Ronald M. Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378.

Kimmo Koskenniemi. 1983. *Two-level Morphology: A General Computational Framework for Word-Form Recognition and Production.* The University of Helsinki, Helsinki, Finnland.

Chris Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachusetts.